Published online 25 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D601–D605***
doi:10.1093/nar/gkq1198

PSSRdb: a relational database of polymorphic simple sequence repeats extracted from prokaryotic genomes

Pankaj Kumar, Pasumarthy S. Chaitanya and Hampapathalu A. Nagarajaram*

Laboratory of Computational Biology, Centre for DNA Fingerprinting and Diagnostics (CDFD), Bulding 7, Gruhakalpa, Nampally, Hyderabad 500 001, India

Received August 17, 2010; Revised October 8, 2010; Accepted November 7, 2010

ABSTRACT

PSSRdb (Polymorphic Simple Sequence Repeats database) (http://www.cdfd.org.in/PSSRdb/) is a relational database of polymorphic simple sequence repeats (PSSRs) extracted from 85 different species of prokaryotes. Simple sequence repeats (SSRs) are the tandem repeats of nucleotide motifs of the sizes 1–6 bp and are highly polymorphic. SSR mutations in and around coding regions affect transcription and translation of genes. Such changes underpin phase variations and antigenic variations seen in some bacteria. Although SSR-mediated phase variation and antigenic variations have been well-studied in some bacteria there seems a lot of other species of prokaryotes yet to be investigated for SSR mediated adaptive and other evolutionary advantages. As a part of our on-going studies on SSR polymorphism in prokaryotes we compared the genome sequences of various strains and isolates available for 85 different species of prokaryotes and extracted a number of SSRs showing length variations and created a relational database called PSSRdb. This database gives useful information such as location of PSSRs in genomes, length variation across genomes, the regions harboring PSSRs, etc. The information provided in this database is very useful for further research and analysis of SSRs in prokaryotes.

INTRODUCTION

Simple sequence repeats (SSRs), also known as microsatellites, are the repetitive nucleotide sequences ubiquitously present in all the known genomes (1–9). These sequences characteristically comprise of mono to hexa nucleotide repeats that are arranged in tandem. SSRs undergo high rates of insertion and deletion (INDEL) mutations of their repeat units as a consequence of slipped mispairing of the nascent and the template strands during replication and hence exhibit high polymorphism (10,11). The INDEL mutations of repeat units in SSRs occurs at high frequencies ranging from 10^{-6} to 10^{-2} per generation, which is much higher than base substitution rates (6,11–13). Mutations in SSRs have different effects depending on the location of SSRs relative to the organization of genes (6,14). SSRs that are located far from coding regions may evolve neutrally and have no effect on structure and function of genes. On the other hand mutations of SSRs either in the coding regions or near the regulatory regions of genes could produce considerable effects on translation or transcription of genes. Furthermore, the severity of the effect in the coding regions depends on the repeat type and the repeat location (11). Polymorphic SSRs of repeating motif length 3 or 6 nt in the coding regions of genome bring out in-frame mutations which translate into insertion or deletion of amino acid residues whereas polymorphic SSRs of non-triplet repeats (mono-, di-, tetra- and penta-nucleotide) bring out frame-shift mutations.

When one looks into abundance and length distribution of SSRs in genomes it gives an impression that SSRs are suppressed in prokaryotic genomes as compared to eukaryotic genomes (9). Nonetheless, some SSRs do show polymorphism and such SSRs have been known to render beneficial effects to prokaryotes [reviewed in (6,8,14)]. The well-documented effects have been the SSR mediated phase variation and antigenic variation which have been well-exploited by many pathogens to evade challenges offered by host immune systems and these have been studied in some bacteria (15).

*To whom correspondence should be addressed. Tel: +91 40 2474 9367; Fax: +91 40 2474 9448; Email: han@cdfd.org.in

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Our group has been analyzing polymorphic SSRs in known prokaryotic genomes and trying to understand evolution of pathogens mediated by SSRs. During the course of our studies, we identified and extracted SSRs which show length variation among different strains and isolates available for 85 different prokaryotic species. All the data pertaining to these polymorphic SSRs (PSSRs) have further been compiled in the form of a relational database called PSSRdb. The present communication gives the details of this database.

EXTRACTION OF THE DATA PERTAINING TO PSSRS

The complete genome sequences of various species with a minimum of two strains were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Extraction of PSSRs was done by an in-house developed tool called PSSRFinder (Kumar, P. and Nagarajaram, H.A., unpublished data) whose workflow is shown in Figure 1. Essentially, PSSRFinder runs BLASTN (16) to identify equivalent SSRs (SSRs having very similar/identical flanking sequences of lengths of at least 50 bp) among all the genomes available for a species.Some essential details of the method are given below:

(i) Identification of SSRs from given genomes using SSRF (17) which reports SSR motif, motif repeat counts, co-ordinate of SSR tract in the genome and its location relative to coding and non-coding regions.
(ii) Identification of equivalent SSRs along with their conserved flanking segments among various strains and isolates by using BLASTN searches with the following set of parameters: E-value $\leq 10^{-20}$; X drop-off value for final gapped alignment = 1000; and repeat masking filter = off.
(iii) Identification of PSSRs by comparing tract lengths of equivalent SSRs found in all the given genomes. If the equivalent polymorphic SSRs are part of non-coding regions in all the genomes it is annotated as non-coding PSSR. If it is found as a part of a coding region even in one of the genomes then the PSSR is referred to as coding PSSR.

STRUCTURE OF THE DATABASE

PSSRdb has been developed using MySql (www.mysql.com). PSSRs found in coding and non-coding regions are separately stored in two different logically connected databases. Both the coding and non-coding databases contain 357 tables each of which contains useful information pertaining to PSSRs viz., motif types, repeat copy numbers of SSRs, genomic location of SSRs and information pertaining to the coding regions harboring or flanking the PSSRs. The details of the structure of the relational tables in the coding and non-coding PSSR databases are given in Tables 1 and 2, respectively.

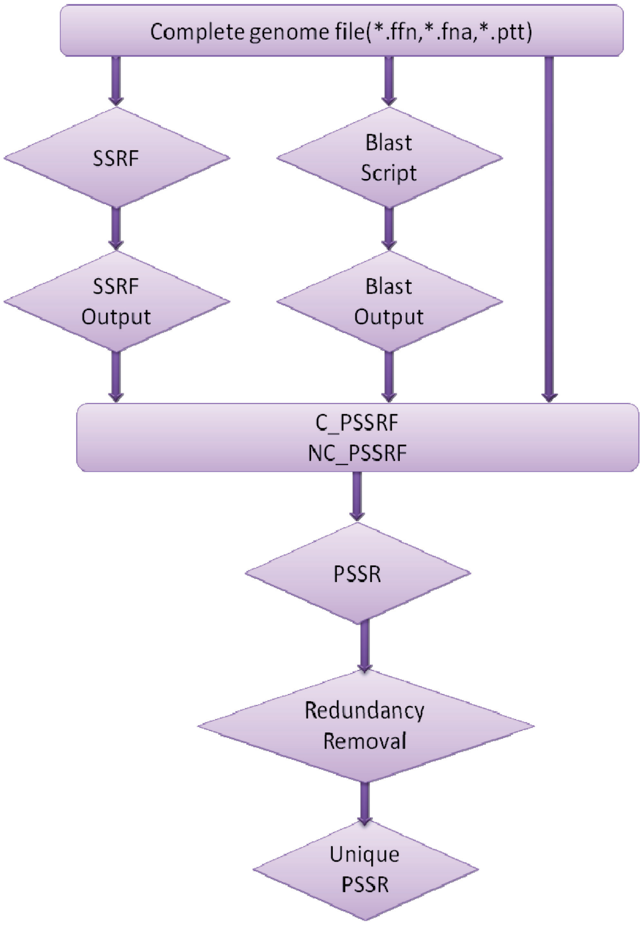


Figure 1. Schematic representation of PSSRFinder. C_PSSRF and NC_PSSRF are the two PERL programs which parse coding and non-coding PSSRs respectively from the BLAST output.

OVERVIEW OF THE DATABASE AND ITS USAGE FOR DATA EXTRACTION

The Database overview is shown in Figure 2. The main page of the database contains a pull down menu containing the names of all the 85 species. Once a selection is made for a species the page is updated with the list of all the available strains belonging to the selected species. One can select two or more of the enlisted strains to query for PSSRs found in those selected set of strains. A separate option is provided to query for PSSRs found in the coding regions and the non-coding regions. A query leads to a page which gives the number of PSSRs found in the selected species. The numbers are clickable links and when clicked display pages containing the detailed information pertaining to the corresponding PSSRs. The displayed information includes the sequence of the repeat motif, its genomic location and the details of the regions harboring that repeat motif. In this page, hyperlinks are also provided to each of the listed PSSRs to design primers using PRIMER3 (14). The coding regions harboring or flanking the PSSRs are also hyperlinked to their respective annotations available at NCBI site (http://www.ncbi.nlm.nih.gov/).

As mentioned earlier, PSSRs stored in PSSRdb have been identified species-wise and these correspond to

Table 1. Structure of MySQL table which is used for storing coding PSSR information

Information	Field	Type	Null	Key	Default	Extra
PSSR number	P_n	int(11)	No	PRI	NULL	auto_increment
Strain name	Strn	varchar(90)	YES		NULL	
PSSR	mf	varchar(8)	YES		NULL	
Repeat length	rpt	int(11)	YES		NULL	
Start of repeat	strt_rpt	varchar(20)	YES		NULL	
End of repeat	end_rpt	varchar(20)	YES		NULL	
Mutation point	mut_pnt	varchar(20)	YES		NULL	
Sequence	seq	varchar(50)	YES		NULL	
Strand type	strnd_type	varchar(5)	YES		NULL	
Protein length	prtn_len	bigint(20)	YES		NULL	
Protein ID	prtn_id	varchar(20)	YES		NULL	
ORF	orf_name	varchar(20)	YES		NULL	
Protein function	prtn_func	varchar(150)	YES		NULL	
DNA sequence of length 400 nucleotides	seq_link	varchar(550)	YES		NULL	

Table 2. Structure of MySQL table which is used for storing non-coding PSSR information

Information	Field	Type	Null	Key	Default	Extra
PSSR number	P_n	int(11)	NO	PRI	NULL	auto_increment
Strain name	Strn	varchar(90)	YES		NULL	
PSSR	mf	varchar(8)	YES		NULL	
Repeat length	rpt	int(11)	YES		NULL	
Start of repeat	s_rpt	varchar(20)	YES		NULL	
End of repeat	e_rpt	varchar(20)	YES		NULL	
Mutation point	mut_pnt	varchar(20)	YES		NULL	
Sequence	seq	varchar(50)	YES		NULL	
Distance from left ORF	L_D	varchar(10)	YES		NULL	
Left strand type	U_S_T	varchar(5)	YES		NULL	
Left protein length	U_P_L	bigint(20)	YES		NULL	
Left protein ID	U_P_I	varchar(20)	YES		NULL	
Left ORF	U_orf	varchar(20)	YES		NULL	
Distance from right ORF	R_D	varchar(10)	YES		NULL	
Right strand type	D_S_T	varchar(5)	YES		NULL	
Right protein length	D_P_L	bigint(20)	YES		NULL	
Right protein ID	D_P_I	varchar(20)	YES		NULL	
Right ORF	D_orf	varchar(20)	YES		NULL	
DNA sequence of 400 nucleotide length	seq_link	varchar(550)	YES		NULL	

those SSRs which show length variation among different strains and isolates available for each of the 85 species. In this respect, we would like to sound a word of caution. Although all the prokaryotic genomes have >10× coverage, some sequencing or assembly mistakes cannot be completely ruled out. Some of SSRs may get qualified as PSSRs as a consequence of sequencing errors or due to mistakes committed during assembly of genome sequences. It is very difficult to identify such artifacts. Nonetheless, we believe the data represented in PSSRdb makes a good starting point for further exploratory investigations on SSR polymorphism in prokaryotes.

The identification of PSSRs in a species has a very good advantage. Depending upon the region of occurrence it could have different potential application. The strain specific PSSR (SSR length varies only in one strain) could be used for the identification of that strain and is of importance in making diagnostic kits. The genes harboring PSSRs form good candidates to study the functional role of genes in pathogenesis and virulence.

FUTURE DIRECTION

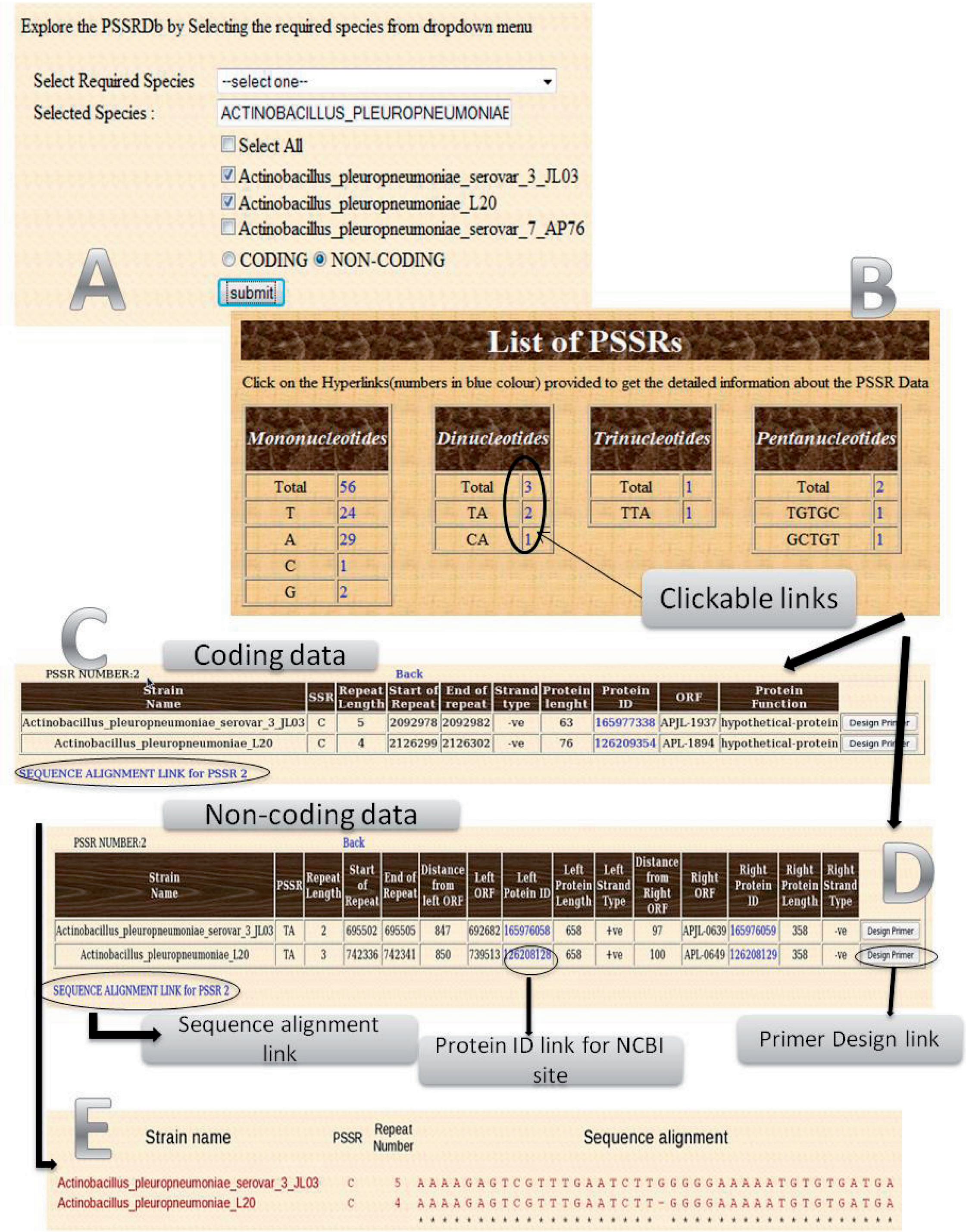
A hyper link will be provided to query for the multiple sequence alignment of the PSSRs along with their flanking regions.So that user can select the number of base pairs from upstream and downstream sequence and will do the multiple sequence alignment on fly. The database will be regularly updated as and when whole genome sequences of new prokaryotes become available.

ACKNOWLEDGEMENTS

P.K. acknowledges Senior Research Fellowship (SRF) from Council of Scientific and Industrial Research (CSIR), India.

FUNDING

The work as well as the publication costs were supported by the Core fund of Centre for DNA Fingerprinting and Diagnostics (CDFD).



Explore the PSSRDb by Selecting the required species from dropdown menu

Select Required Species: --select one--

Selected Species : ACTINOBACILLUS_PLEUROPNEUMONIAE

☐ Select All
☑ Actinobacillus_pleuropneumoniae_serovar_3_JL03
☑ Actinobacillus_pleuropneumoniae_L20
☐ Actinobacillus_pleuropneumoniae_serovar_7_AP76

○ CODING ● NON-CODING

submit

List of PSSRs

Click on the Hyperlinks(numbers in blue colour) provided to get the detailed information about the PSSR Data

Mononucleotides	
Total	56
T	24
A	29
C	1
G	2

Dinucleotides	
Total	3
TA	2
CA	1

Trinucleotides	
Total	1
TTA	1

Pentanucleotides	
Total	2
TGTGC	1
GCTGT	1

PSSR NUMBER:2 Back

Strain Name	SSR	Repeat Length	Start of Repeat	End of repeat	Strand type	Protein lenght	Protein ID	ORF	Protein Function	
Actinobacillus_pleuropneumoniae_serovar_3_JL03	C	5	2092978	2092982	-ve	63	165977338	APJL-1937	hypothetical-protein	Design Primer
Actinobacillus_pleuropneumoniae_L20	C	4	2126299	2126302	-ve	76	126209354	APL-1894	hypothetical-protein	Design Primer

SEQUENCE ALIGNMENT LINK for PSSR 2

PSSR NUMBER:2 Back

Strain Name	PSSR	Repeat Length	Start of Repeat	End of Repeat	Distance from left ORF	Left ORF	Left Potein ID	Left Protein Length	Left Strand Type	Distance from Right ORF	Right ORF	Right Protein ID	Right Protein Length	Right Strand Type	
Actinobacillus_pleuropneumoniae_serovar_3_JL03	TA	2	695502	695505	847	692682	165976058	658	+ve	97	APJL-0639	165976059	358	-ve	Design Primer
Actinobacillus_pleuropneumoniae_L20	TA	3	742336	742341	850	739513	126208128	658	+ve	100	APL-0649	126208129	358	-ve	Design Primer

SEQUENCE ALIGNMENT LINK for PSSR 2

```
Strain name                                     PSSR  Repeat  Sequence alignment
                                                      Number
Actinobacillus_pleuropneumoniae_serovar_3_JL03  C     5       AAAAGAGTCGTTTGAATCTTGGGGAAAAATGTGTGATGA
Actinobacillus_pleuropneumoniae_L20             C     4       AAAAGAGTCGTTTGAATCTT-GGGAAAAATGTGTGATGA
                                                              ******************** ******************
```

Figure 2. Overview of PSSRdb shown using screen-shots of various pages. (**A**) Main page containing species name which can be selected; (**B**) PSSRs found in the selected species; (**C**) Table containing the useful details of the selected coding PSSRs found in the selected species; (**D**) Table containing the useful details of the selected non-coding PSSRs found in the selected species; (**E**) Sequence alignment of a selected PSSR (in this case G tract).

Conflict of interest statement. None declared.

REFERENCES

1. Schlotterer,C. and Tautz,D. (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.*, **20**, 211–215.
2. Tautz,D. (1993) Notes on the definition and nomenclature of tandemly repetitive DNA sequences. *EXS*, **67**, 21–28.
3. Moxon,E.R., Rainey,P.B., Nowak,M.A. and Lenski,R.E. (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.*, **4**, 24–33.
4. Tautz,D. and Schlotterer,C. (1994) Simple sequences. *Curr. Opin. Genet. Dev.*, **4**, 832–837.
5. Schlotterer,C. (1998) Genome evolution: are microsatellites really simple sequences? *Curr. Biol.*, **8**, R132–R134.

6. van Belkum,A., Scherer,S., van Alphen,L. and Verbrugh,H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.*, **62**, 275–293.
7. Buschiazzo,E. and Gemmell,N.J. (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, **28**, 1040–1050.
8. Moxon,R., Bayliss,C. and Hood,D. (2006) Bacterial contingency Loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.*, **40**, 307–333.
9. Mrazek,J., Guo,X. and Shah,A. (2007) Simple sequence repeats in prokaryotic genomes. *Proc. Natl Acad. Sci. USA*, **104**, 8472–8477.
10. Levinson,G. and Gutman,G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.
11. Sreenu,V.B., Kumar,P., Nagaraju,J. and Nagarajaram,H.A. (2006) Microsatellite polymorphism across the M. tuberculosis and M. bovis genomes: implications on genome evolution and plasticity. *BMC Genomics*, **7**, 78.
12. Garcia-Diaz,M. and Kunkel,T.A. (2006) Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem. Sci.*, **31**, 206–214.
13. Kunkel,T.A. (2004) DNA replication fidelity. *J. Biol. Chem.*, **279**, 16895–16898.
14. v'an der Woude,M.W. and Baumler,A.J. (2004) Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.*, **17**, 581–611.
15. Brunham,R.C., Plummer,F.A. and Stephens,R.S. (1993) Bacterial antigenic variation, host immune response, and pathogen-host coevolution. *Infect. Immun.*, **61**, 2273–2276.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Sreenu,V.B., Ranjitkumar,G., Swaminathan,S., Priya,S., Bose,B., Pavan,M.N., Thanu,G., Nagaraju,J. and Nagarajaram,H.A. (2003) MICAS: a fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. *Appl. Bioinformatics*, **2**, 165–168.